التحليل الصرفي لنصوص اللغة العربية الحديثة والكلاسيكية

مجدي صوالحه ، إرك أتول

العنوان البريدي: School of Computing, University of Leeds, Leeds, LS2 9JT, UK.

البريد الالكتروني: scmss@leeds.ac.uk ، csc6ea@leeds.ac.uk

الخلاصة. في هذا البحث سنقوم بعرض تجربة في توظيف معايير وأدوات التحليل الصرفي للغة العربية، بالإضافة إلى المعلومات اللغوية المستخرجة من كتب قواعد اللغة العربية التقليدية، لتدوير وإغناء المصادر اللغوية المتاحة بتحليل صرفيٍّ دقيق يحتوي على جميع الخصائص الصرفية للكلمة المحللة، وقد تمَّ استخدام مجموعة العناوين الصرفية (سلمى[1]) كمعيار للتحليل الصرفي في هذه الدراسة، كما تمَّ استخدام المحلل الصرفي (سلمى) [1-3] كأداة للتحليل الصرفي.

إن تدوير المصادر المتاحة هو مبدأ أساسي من مبادئ هندسة البرمجيات، لقد تم استخدام العنونة الصرفية للذخيرة العربية القرآنية[2] [4] لعينة مكونة من حوالي ألف كلمة (سورة العنكبوت – رقم 29)، وبعدها تم تحويل العناوين الصرفية آلياً من مجموعة العناوين الصرفية الخاصة بالذخيرة القرآنية إلى مجموعة العناوين (سلمى) ذات الخصائص الصرفية الدقيقة.

وقد تمت عملية التحويل أو التدوير لهذه العناوين الصرفية بتنفيذ إجراء حاسوبي مكون من خمسة خطوات، وقد برهنت النتائج عن قابلية تدوير وإغناء المصادر المتاحة بتحليل صرفي دقيق منتجةً ذخيرة لغوية معنونة صرفياً بتحليل صرفي مفصل ودقيق.

الكلمات الجوهرية: التحليل الصرفي، اللغة العربية، القرآن الكريم، التدوير، مجموعة العناوين الصرفية (سلمى)، المعيار الذهبي لتقييم المحللات الصرفية لنصوص اللغة العربية.

1 مقدمة

لقد تمت دراسة التحليل الصرفي للغة العربية من قبل العديد من الباحثين، ولقد طبق الباحثون عدة تقنيات لحل المشكلات الصرفية المعقدة للغة العربية، كتقنيات الإعتمادية المعرفية والتعلم الآلي، نريد عنونة الذخيرة العربية صرفياً بأقسام الكلام، ولكن تقييم المحللات الصرفية المتاحة أظهر قصورها وتحديداً: حوالي ربع كلمات النص المحلل لم يتم عنونتها بشكل كامل وصحيح، وذلك لأن العنونة الصرفية لنصوص اللغة العربية أكثر تعقيداً [5]. يستخدم المُجَذِّع لاستخراج جذع الكلمة و جذرها، ولكن هذه المعلومات المستخرجة غير كافية للعنونة الصرفية الكاملة، العنونة الصرفية الآلية لنصوص الغة العربية ينبغي فيها أن يتم تقسيم الكلمة إلى خمسة أجزاء؛ لواصق في بداية

---

[1] مجموعة عناوين الخصائص الصرفية (سلمى) http://www.comp.leeds.ac.uk/sawalha/tagset.html

[2] الذخيرة العربية القرآنية http://corpus.quran.com/

الكلمة، و زوائد في بداية الكلمة، وجذع الكلمة أو جذرها، وزوائد في نهاية الكلمة، ولواصق في نهاية الكلمة، ويجب على المحلل الصرفي إضافة المعلومات اللغوية لكل جزء من أجزاء الكلمة، وبالنتيجة بدلاً من عنوان صرفي واحد للكلمة، نحن بحاجة إلى عنوان صرفي لكل جزء (ومن الممكن إضافة أكثر من عنوان صرفي لهذه الأجزاء إذا احتوت على أكثر من سابقة أو لاحقة)، وللغة العربية العديد من الخصائص الصرفية والنحوية والتي تشمل الإسناد والمذكر والمؤنث والحالة الإعرابية للاسم والفعل ...الخ [6]، وإن مجموعة العناوين الصرفية الدقيقة تُفضل على غيرها، وتساعد المعلومات الصرفية الإضافية على إزالة الغموض للعنوان الصرفي الرئيسي للكلمة [7] [2].

للتحليل الصرفي عدة تطبيقات في مجال معالجة اللغات الطبيعية، ففي البحث في شبكة الإنترنت باستخدام لغة ذات قواعد صرفية معقدة، يسمح التحليل الصرفي بالبحث عن مشتقات الكلمة المراد البحث عنها، كما يعطي التحليل الصرفي أهم المعلومات اللغوية لبرنامج العنونة الآلية بأقسام الكلام لتمكينه من اختيار أفضل تحليل صرفي مناسب لسياق معين، ويعتمد بناء المعاجم اللغوية وأنظمة التدقيق الإملائي على محللات صرفية عالية الوثوقية حتى تتمكن من تحديد ترجمة صحيحة للجمل المدخلة، ويعتبر التجذيع جزءاً من التحليل الصرفي، ويعتمد نظام البحث في جوجل على برنامج التجذيع للبحث عن جميع مشتقات الكلمة، ويستخدم برنامج التجذيع في النماذج الإحصائية اللغوية كالمترجمات النَّصية، وتطبيقات التعرف على الكلام، وأيظاً تعتبر برامج تجذيع الكلمات من الأدوات الأساسية في التطبيقات المعجمية وبناء الذخائر اللغوية [8].

1.1 التحليل الصرفي لنصوص اللغة العربية

اللغة العربية هي لغة صرفية ذات قواعد معقدة، مما يجعل مهمة تحليل النصوص العربية غاية في الصعوبة، إن الأسلوب الصرفي المعتمد على الجذر والميزان الصرفي والنظام المعقد لإشتقاق الكلمات من جذورها حسب أوزان معينة، والتي من خلالها يمكن اشتقاق المئات من الكلمات من جذر واحد، وإن إضافة اللواصق والزوائد في بداية الكلمة ونهايتها يزيد بشكل كبير من عدد مشتقات الكلمة، وكذلك فإن الغموض في النصوص العربية من التحديات الرئيسية للمعالجة الآلية، ويزداد الغموض بإختفاء التشكيل في معظم النصوص العربية، كما يزداد أيضاً بالتشابه بين حروف اللواصق والزوائد من جهة، وحروف الكلمة الأصلية من جهة أخرى، خاصةً إذا كان حرف أو حرفان من أحرف الكلمة الأصلية أحرف علة، ولذلك فإن التحليل الصرفي لنصوص اللغة العربية مهمة صعبة وله تأثير مباشر على التطبيقات اللغوية العالية المستوى كالعنونة الآلية لأقسام الكلام والتحليل النحوي [2].

تعتبر اللواصق والزوائد في اللغة العربية إنتاجية، ولذلك فإنه من غير الممكن اعتماد التحليل الصرفي على قاموس شامل يحتوي على جميع أشكال الكلمة ومشتقاتها من خلال البحث عن شكل معين للكلمة في هذا القاموس [9]، وعليه ينبغي أن يكون التحليل الصرفي آلياً، فعلى سبيل المثال، أحد التحليلات الممكنة لكلمة (بِوَالِدَيْهِ) مكون من أربعة مقاطع هي: حرف الجر (ب)، والجذع الاسمي (والد)، وحرف التثنية (ي)، والضمير المتصل (ه)، وهي لواصق إنتاجية.

2 سلمى – المحلل الصرفي ومجموعة عناوين الخصائص الصرفية الدقيقة

في هذا البحث نعرض بعض معايير وأدوات التحليل الصرفي للغة العربية التي تم تطويرها، والتي يمكن استخدامها لتدوير وإغناء المصادر اللغوية المتاحة بتحليل صرفيٍّ دقيق. نعرض في الجزء 2.1 المحلل الصرفي (سلمى) كأداة للتحليل الصرفي ، وفي الجزء 2.2 مجموعة عناوين الخصائص الصرفية الدقيقة (سلمى) كمعيار للتحليل الصرفي.

2.1 سلمى – المحلل الصرفي لنصوص اللغة العربية

يعتمد المحلل الصرفي (سلمى)[3] [1-3] على المعلومات اللغوية للغة العربية وعلى المعجم واسع التغطية والذي أُنشأ اعتماداً على تحليل 23 معجماً عربياً تقليدياً [10]، كما يعتمد أيضاً على قوائم شاملة للزوائد والسوابق واللواحق والأوزان تم استخراجها من الكتب القيمة للنحو والصرف وقواعد اللغة العربية، ويعتمد المحلل الصرفي على هذه القوائم لتحليل الكلمات، كما يعتمد على خوارزمية جديدة لتحديد وزن الكلمة الصحيح إذا احتوت هذه الكلمات على حروف علَّة أو همزة أو إعلال أو إقلاب، ويقوم المحلل الصرفي (سلمى) بتقطيع الكلمة العربية إلى لواصق في بداية الكلمة، وزوائد في بداية الكلمة، وجذع الكلمة او جذرها، وزوائد في نهاية الكلمة، ولواصق في نهاية الكلمة، ويعين المحلل الصرفي (سلمى) عنواناً صرفيّاً لكل مقطع من مقاطع الكلمة المحللة.

2.2 سلمى – مجموعة عناوين الخصائص الصرفية الدقيقة[4]

قسَّم علماء اللغة العربية وقواعدها الكلمة إلى ثلاثة أقسام رئيسية هي الاسم والفعل والحرف، وتم وصف وتفصيل وتحديد خصائص هذه الأقسام بدقة، كما تمَّ تحديد الخصائص اللغوية للكلمة كالجنس والعدد والإسناد والحالة الإعرابية للاسم أو الفعل ، والمعرفة والنكرة، وبناء الفعل للمعلوم أو المجهول، والتَّوكيد، والفعل اللازم أو المتعدي، ومن الخصائص الأخرى التي تمَّ تحديدها للكلمة والتي تصف بنيتها، المجرَّد والمزيد، وعدد حروف الكلمة الأصلية، وتركيب أحرف الفعل الثلاثي من حيث الصحة والإعلال.

اعتماداً على هذه الخصائص اللغوية تمَّ تصميم مجموعة عناوين الخصائص الصرفية (سلمى) حيث يتكون العنوان الصرفي الدقيق من 22 رمزاً، كل رمز يمثل قيمة أو متغير ينتمي إلى إحدى الخصائص الصرفية، ويعدُّ موقع الرمز في العنوان مهماً في تحديد هذه الخاصية، وتُمثَّل هذه القيم أو المتغيرات برمز واحد من حروف اللغة الإنجليزية الصغيرة، فمثلا الرمز (v) في الموقع الأول من العنوان يرمز إلى الفعل، والرمز (n) في الموقع الثاني يرمز إلى اسم العلم، ويمثل الجنس في الموقع السابع من العنوان حيث يرمز الحرف (m) الى المذكر والحرف (f) الى المؤنث، وإذا كانت الخاصية اللغوية غير متوافقة مع الكلمة فالرمز (-) (الشرطة) يمثلها، بينما يستخدم الرمز (؟) (علامة السؤال) لترمز إلى أن الخاصية اللغوية تنطبق على الكلمة ولكن غير محددة.

---

[3] المحلل الصرفي (سلمى) – (SALMA Tagger (Sawalha Atwell Leeds Arabic Morphological Analysis – Tagger)

[4] مجموعة العناوين (سلمى) - (SALMA - Tag Set (Sawalha Atwell Leeds Morphological Analysis - Tag Set))

3 تطوير المعيار الذهبي الدقيق للتحليل الصرفي

تستخدم المعايير الذهبية لتقييم وقياس دقة الأنظمة المحوسبة، كما يمكن استخدامها للمقارنة بين عدَّة أنظمة أو خوارزميات طورت لحل مشكلة معيَّنة، وتُظهر المعايير الحالات التي تنجح أو تفشل الأنظمة المُقيَّمة بتحديد التحليل المناسب للمدخلات، وتستخدم المعايير الذهبية لإيجاد أوجه الشبه أو الإختلاف في نتائج التحليل مبينة الحالات التي تتفق عليها والتي تختلف فيها الأنظمة المحوسبة[2].

إن تدوير المصادر المتاحة هو مبدأ أساسي من مبادئ هندسة البرمجيات ، ومن المصادر الجديدة والمتاحة الذخيرة اللعربية القرآنية [4] والتي تحتوي على طبقات متعددة من التحليل اللغوي والذي يشمل التحليل الصرفي لكلمات القرآن الكريم، لقد تم استخدام العنونة الصرفية للذخيرة العربية القرآنية لعينة مكونة من حوالي ألف كلمة (سورة العنكبوت – رقم 29)، وبعدها تم تحويل العناوين الصرفية آلياً من مجموعة العناوين الصرفية الخاصة بالذخيرة القرآنية إلى مجموعة العناوين (سلمى) ذات الخصائص الصرفية الدقيقة.

3.1 التحويل من مجموعة العناوين الصرفية للذخيرة العربية القرآنية إلى مجموعة عناوين الخصائص الصرفية الدقيقة (سلمى)

تمت عملية تحويل العناوين الصرفية المستخدمة في الذخيرة العربية القرآنية، إلى مجموعة عناوين الخصائص الصرفية الدقيقة (سلمى)، بإتباع إجراء حاسوبي مكون من خمسة خطوات، نعرض في هذا الجزء من البحث وصف لعملية التحويل، من خلاله نعرض التحديات التي واجهتنا في عملية التحويل، كما توضح الأمثلة المرفقة عينات من التحليل الصرفي لأجزاء من عينة الإختبار تمت عنونتها باستخدام مجموعة عناوين الخصائص الصرفية الدقيقة (سلمى).

3.1.1 تحويل النص القرآني من الرسم العثماني إلى الرسم الإملائي

استخدمت الذخيرة العربية القرآنية، نصوص القرآن الكريم المكتوبة بالرسم العثماني، ولكن معظم أدوات المعالجة الآلية للغة العربية تتعامل فقط مع نصوص اللغة العربية الحديثة والمكتوبة بالرسم الإملائي، وتحتاج هذه الأدوات لبعض التعديلات لتمكينها من التعامل مع النصوص العربية الكلاسيكية وخاصة المكتوبة بالرسم العثماني، ولحسن الحظ؛ توجد نسخة من القرآن الكريم مكتوبة بالرسم الإملائي، وقد تمت عملية التحويل من الرسم العثماني إلى الرسم الإملائي حسب علاقة واحد لواحد لربط كلمات القرآن الكريم إلا لبعض الحالات بسبب الإختلاف في كتابة الكلمات بين الرسمين، فمثلاً؛ حرف النداء (يا) يكتب موصولاً بالمنادى في الرسم العثماني ويكتب منفصلاً في الرسم الإملائي، وعلية فإن الذخيرة العربية القرآنية تحتوي على 77430 كلمة حسب الرسم العثماني تم تحويلها إلى ما يقابلها بالرسم الإملائي وعددها 77797 كلمة (الشكل 1).

الرسم العثماني	الرسم الإملائي	الرسم العثماني	الرسم الإملائي
يَٰمُوسَىٰ	يَا مُوسَى	وَأَلَّوِ	وَأَنْ لَوِ
يَٰأَهْلَ	يَا أَهْلَ	يَبْنَؤُمَّ	يَا ابْنَ أَمَّ
يَٰلَيْتَنِى	يَا لَيْتَنِي	هَٰأَنتُمْ	هَا أَنْتُمْ

الشكل .1 أمثلة على الإختلافات الكتابية بين الرسمين العثماني والإملائي

3.1.2 تقسيم كلمات النص القرآني إلى مقاطعها الصرفية

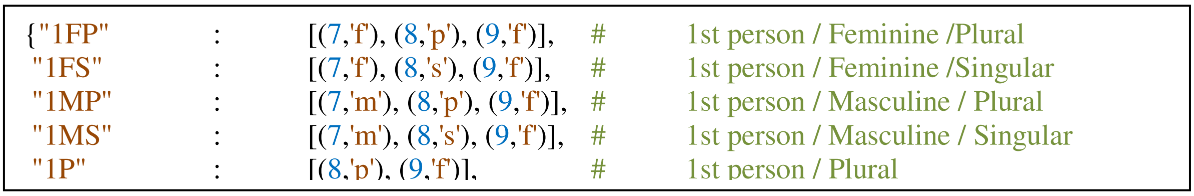
تقسيم كلمات النص القرآني في عينة الاختبار إلى مقاطعها الصرفية لم يكن بالأمر السهل، وذلك لأن كلمات القرآن الكريم في الذخيرة العربية القرآنية غير مقسمة، والإشارة الوحيدة على المقاطع الصرفية للكلمة توجد في العنوان الصرفي لتلك الكلمة، حيث تم فصل أجزاء العنوان الصرفي لمقاطع الكلمة بإشارة (+)، ومنها تم تقسيم العنوان الصرفي والكلمة آلياً وتعيين أجزاء العنوان الصرفي المناسب لمقاطع الكلمة (الشكل 2).

أَفَبِٱلْبَٰطِلِ	أَفَبِالْبَاطِلِ	A:INTG+ f:REM+ bi+ Al+ POS:N ACT PCPL LEM:ba`Til ROOT:bTl M GEN	
أ	أ	A:INTG	المقطع (1)
فَ	فَ	f:REM	المقطع (2)
بِ	بِ	bi	المقطع (3)
ٱلْ	الْ	Al	المقطع (4)
بَٰطِلِ	بَاطِلِ	POS:N ACT PCPL LEM:ba`Til ROOT:bTl M GEN	المقطع (5)

الشكل .2 تقسيم الكلمة إلى مقاطعها وتعيين الجزء المناسب لها من العنوان الصرفي.

3.1.3 التحويل المباشر بين مجموعتي العناوين الصرفية

تم بناء قاموس يربط مباشرة بين رموز العناوين الصرفية المستخدمة في الذخيرة العربية القرآنية وما يقابلها من رموز مجموعة عناوين الخصائص الصرفية الدقيقة (سلمى)، وإحتوى هذا القاموس على 158 مدخلة، تربط الرمز المراد استعماله بموقعه في العنوان الصرفي لمجموعة (سلمى)، ويوضح الشكل 3 جزءاً من هذا القاموس والذي تم استخدامه من قبل برنامج التحويل.



```
{"1FP"     :     [(7,'f'), (8,'p'), (9,'f')],    #     1st person / Feminine /Plural
 "1FS"     :     [(7,'f'), (8,'s'), (9,'f')],    #     1st person / Feminine /Singular
 "1MP"     :     [(7,'m'), (8,'p'), (9,'f')],    #     1st person / Masculine / Plural
 "1MS"     :     [(7,'m'), (8,'s'), (9,'f')],    #     1st person / Masculine / Singular
 "1P"      :     [(8,'p'), (9,'f')],             #     1st person / Plural
```

الشكل 3. جزء من قاموس التحويل المباشر بين رموز مجموعتي العناوين الصرفية

3.1.4 إضافة الخصائص الصرفية غير الموجودة في العناوين الصرفية للذخيرة العربية القرآنية

تم تصميم إجراءات حاسوبية لمعرفة قيمة الخصائص الصرفية غير المتوفرة في العناوين الصرفية للذخيرة العربية القرآنية، وقد تم الاعتماد على قواعد اللغة العربية المستخرجة من كتب قواعد الصرف العربي لبناء نظام خبير يحلل مقاطع الكلمات ويعين قيمة الخاصية الصرفية لكل مقطع بدقة، ويوضح الشكل 5 هذه الخصائص من خلال جزء من عينة الاختبار.

3.1.5 تصحيح العناوين الصرفية

تمت عملية التصحيح يدوياً من قبل متخصص في اللغة العربية، وإن ناتج عملية التصحيح هذه أفضى إلى إصدار المعيار الذهبي القرآني[5] لعينة مكونة من حوالي ألف كلمة (سورة العنكبوت –

---

[5] المعيار الذهبي القرآني الدقيق http://www.comp.leeds.ac.uk/sawalha/goldstandard.html

رقم 29)، ويوضح الشكل 5 جزءاً من عينة الإختبار بعد تطبيق الخطوات الثلاثة الأولى وبعد الخطوة الرابعه وبعد تصحيح العناوين الصرفية يدوياً في الخطوة الخامسة.

4 التقييم والنتائج

تضمنت عملية تصحيح العناوين الصرفية التي تم تحويلها آلياً، تصحيح قيم الخصائص الصرفية جميعها لكل مقطع صرفي من مقاطع كلمات عينة الاختبار، وفيها تم تحديد قابلية الخاصية الصرفية لذلك المقطع من الكلمة المحللة، فإذا كانت الخاصية الصرفية هي إحدى الخصائص الصرفية للمقطع المحلل يتم تدقيق قيمتها وتصحيحها إذا كانت خاطئة، وإذا لم توافق الخاصية الصرفية ذلك المقطع يتم التأكد من وجود (-) كقيمة لتلك الخاصية الصرفية، ويوضح الشكل 4 قيم الشواهد التي تم تعيينها خلال عملية التصحيح.

- TN: True and not applicable; cases were negative and predicted negative.
- TP: True and applicable; cases were applicable and predicted correctly.
- FN: False and not applicable; cases were not applicable and predicted applicable.
- FP: False and applicable; cases were applicable and predicted not applicable.

الشكل 4. المشاهدات التي تم تعيينها خلال عملية التصحيح

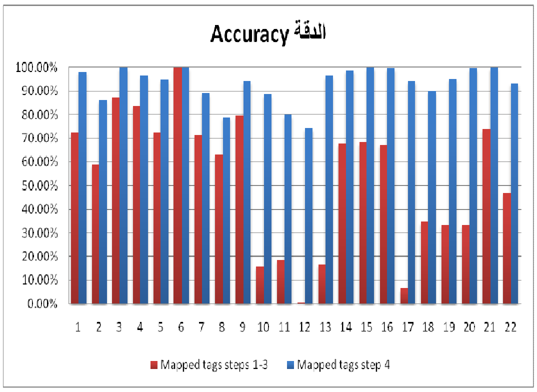
وقد اعتمدت قيم هذه المشاهدات كأساس لمقاييس تقييم نتائج تحويل العناوين الصرفية وإغناؤها، وقد استعمل المعيار الذهبي الدقيق الناتج من عملية تصحيح العناوين الصرفية كمعيار لقياس دقة خوارزمية التحويل، حيث تمت مقارنة العناوين الصرفية الناتجة من الخطوات الثلاثة الأولى والعناوين الصرفية الناتجة من بعد تطبيق الخطوة الرابعة بالعناوين الصرفية في المعيار الذهبي الدقيق، وقد تم استخدام ثلاثة مقاييس لتقييم النتائج هي: الدقة، والضبط، والتذكر، وتُعَرَّفُ الدقةُ بنسبة التوقعات الصحيحة ، وتحسب وفق المعادلة (الدقة = (TN+TP) / عدد مقاطع الكلمات المحللة)، ويُعَرَّفُ التذكرُ بأنه نسبة الحالات التي تم تحويلها بشكل صحيح من مجموع الحالات التي تحولها، وتحسب وفق المعادلة (التذكر = TP / (TP + FN))، ويُعَرَّفُ الضبطُ بأنه نسبة الحالات الصحيحة التي تكون فيها الخاصية الصرفية قابلة للتحويل إلى مجموع الحالات التي تكون فيها الخاصية الصرفية قابلة للتحويل، وتحسب بالمعادلة (الضبط = TP / مجموع الحالات القابلة للتحويل). يعرض الجدول (1) قيم مقاييس التقييم الثلاثة التي استعملت لتقييم نتائج تحويل العناوين الصرفية بعد الخطوات الثلاثة الأولى وبعد تطبيق الخطوة الرابعة لكل خاصية صرفية من خصائص العنوان الصرفي الإثنتين والعشرين، كما تعرض الأشكال 6 و 7 و8 رسوماً توضيحيةً تبين قيم المقاييس الثلاثة ومقدار الزيادة فيها بعد تطبيق الخطوة الرابعة والتي استخدمت النظام الخبير المستمد من برمجة القواعد الصرفية المستخرجة من كتب قواعد اللغة العربية التقليدية.

وكانت نسبة تصحيح العنوان الصرفي لمقاطع الكلمات في عينة الإختبار حوالي 53.5%، بعض التصحيح الذي تم للعناوين الصرفية كان بسيطاً جداً كاستبدال "؟" ب "-"، وكانت نسبة التصحيح للخصائص الصرفية منفردة كالتالي: 2.01% لأقسام الكلام الرئيسية، وما بين 3% و 15% للخصائص الصرفية المعنونة في الذخيرة العربية القرآنية، وما بين 2% و 24% للخصائص

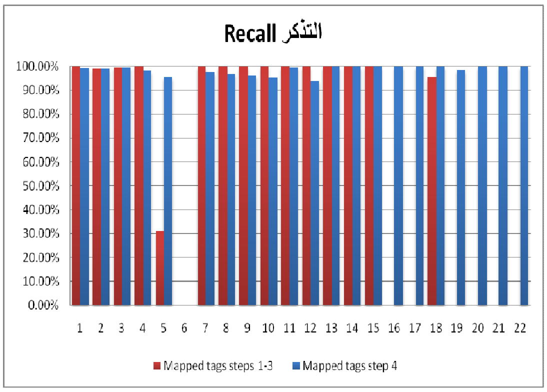
الصرفية غير المعنونة في الذخيرة العربية القرآنية والتي تم عنونتها آلياً، ونظراً لاستخدام 22 خاصية صرفية لكل عنوان صرفي لمقاطع الكلمات المحللة، والتي تزيد من إحتمالية أخطاء العنونة، تبرهن هذه النتيجة قابلية تدوير وإغناء المصادر المتاحة بالتحليل الصرفي الدقيق والتي تنتج ذخيرة لغوية معنونة صرفياً بتحليل صرفي مفصل ودقيق.

المقطع	العنوان الصرفي في الذخيرة العربية القرآنية	العنوان الصرفي بعد تطبيق الخطوات 1-3	العنوان الصرفي بعد تطبيق الخطوة 4	العنوان الصرفي بعد تصحيحه يدوياً (الخطوة 5)
الم	POS:INL	p--?-----????---?-----	p--?-----s-s----------	p--b-----s-s----------
أ	A:INTG+	p--i-----s------------	p--i-----s------------	p--i-----s------------
حَسِبَ	POS:V PERF 3MS	v-p---mst--?-?-??????-	v-p---msts-f-ambhvsta-	v-p---msts-f-amohvsta-
ال	Al+	r---d-----------------	r---d-----------------	r---d-----------------
نَاسُ	POS:N MP NOM	n?----mp-?n??---????-?	n?----mp-vndd---ndst-s	n#----mj-vndd---hdst-s
أنْ	POS:SUB	p--g-------?----------	p--g-----s-s----------	p--g-----s-s----------
يُ	NULL	r---a-----------------	r---a-----------------	r---a-----------------
تْرَكُ	POS:V IMPF PASS 3MP MOOD:SUBJ	v-c---mptda?-p???????-	v-c---mptdao-pmbhvtta-	v-c---mptdao-pmohvtta-
وا	PRON:3MP	r---r-mptsnw----------	r---r-mptsnw----------	r---r-mpts-s----------
أنْ	POS:SUB	p--g-------?----------	p--g-----s-s----------	p--g-----s-s----------
يَ	NULL	r---a-----------------	r---a-----------------	r---a-----------------
قُولُ	POS:V IMPF 3MP MOOD:SUBJ	v-c---mptda?-????????-	v-c---mptdao-amohvtto-	v-c---mptdao-amohvtto-
وا	PRON:3MP	r---r-mptsnw----------	r---r-mptsnw----------	r---r-mpts-s----------
آمَنَ	POS:V PERF (IV) 1MP	v-p---mpf--?-?-??????-	v-p---mpfs-s-amohvttc-	v-p---mpfs-s-amohvttc-
نَا	PRON:1MP	r---r-xpfs??----------	r---r-xpfs??----------	r---r-xpfs-s----------
وَ	wa+	p--c------------------	p--c------------------	p--c-----s-f---------
هُمْ	POS:PRON 3MP	np----mpt--??---?-----	np----mpts-si---hn---?	np----mpts-si---hn----
لَا	POS:NEG	p--n-------?----------	p--n-----s-s----------	p--n-----s-s----------
يُ	NULL	r---a-----------------	r---a-----------------	r---a----------------
فْتَنُ	POS:V IMPF PASS 3MP	v-c---mpt-??-p???????-	v-c---mptdnn-pmohvtta-	v-c---mptdnn-pmohvtta-
ونَ	PRON:3MP	r---r-mp?snn----------	r---r-mp?snn----------	r---r-mpts-f----------

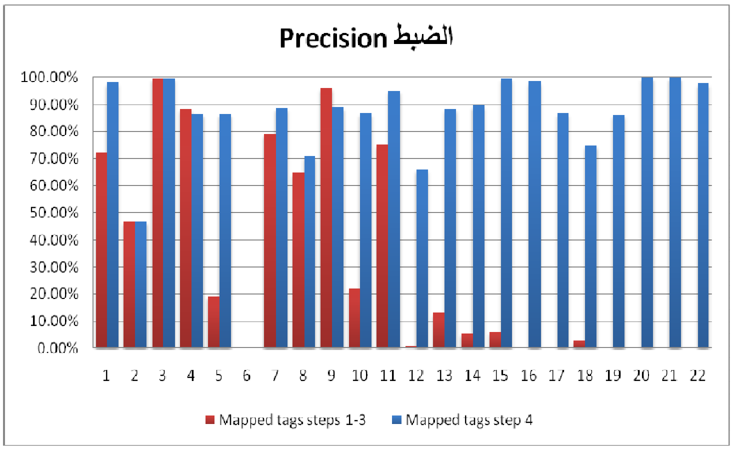
الشكل 5. جزءاً من عينة الإختبار توضح نتائج خطوات تحويل العناوين الصرفية وبناء المعيار الذهبي الدقيق



الشكل 6. دقة تحويل العناوين الصرفية



الشكل 7. التذكر في تحويل العناوين الصرفية



الشكل 8. الضبط في تحويل العناوين الصرفية

الجدول 1. قيم مقاييس التقييم الثلاثة بعد تطبيق الخطوات الثلاثة الأولى و بعد الخطوة الرابعة لكل خاصية صرفية

الخصائص الصرفية		العنوان الصرفي بعد تطبيق الخطوات 3-1			العنوان الصرفي بعد تطبيق الخطوة 4		
		الدقة	التذكر	الضبط	الدقة	التذكر	الضبط
1	أَقسام الكلام الرئيسيَّة	72.30%	100.00%	72.30%	97.99%	99.43%	97.99%
2	أقسام الكلام الفرعيَّة (الاسم)	58.96%	99.16%	46.81%	86.15%	99.16%	46.81%
3	أقسام الكلام الفرعيَّة (الفعل)	87.18%	99.62%	99.62%	99.95%	99.62%	99.62%
4	أقسام الكلام الفرعيَّة (الحرف)	83.73%	100.00%	88.37%	96.24%	98.03%	86.63%
5	أقسام الكلام الفرعيَّة (أخرى)	72.45%	30.84%	19.31%	94.90%	95.50%	86.43%
6	علامات الترقيم	100.00%	-	-	100.00%	-	-
7	المُذكَّر والمُؤَنَّث	71.11%	100.00%	79.11%	89.03%	97.66%	88.72%
8	العدد	63.13%	100.00%	64.82%	79.09%	97.09%	70.91%
9	الاسناد	79.40%	100.00%	96.23%	94.28%	96.11%	89.02%
10	الصَّرف	15.65%	100.00%	22.04%	88.47%	95.30%	86.73%
11	الحالة الإعرابية للاسم أو الفعل	18.54%	100.00%	75.31%	79.71%	99.56%	94.98%
12	علامة الإعراب أو البناء	0.41%	100.00%	0.58%	74.25%	94.20%	66.11%
13	المَعْرِفَة والنَّكِرَة	16.68%	100.00%	12.96%	96.40%	100%	88.46%
14	المَبْني لِلمَعْلُوم و المَبْني لِلمَجْهُول	67.97%	100.00%	5.38%	98.61%	100%	89.62%
15	المُؤكَّد وغيرُ المُؤكَّد	68.07%	100.00%	6.15%	99.95%	100%	99.62%
16	اللازم والمتعدي	67.25%	0.00%	0.00%	99.69%	100%	98.45%
17	العاقل وغير العاقل	6.59%	0.00%	0.00%	94.34%	100%	86.68%
18	التَّصريف	34.65%	95.65%	2.89%	90.11%	99.83%	75.03%
19	المجرَّد والمزيد	33.37%	0.00%	0.00%	95.21%	98.56%	86.19%
20	عَدَد أحْرُف الجَذْر	33.42%	0.00%	0.00%	99.74%	100%	100%
21	بُنية الفعل	73.84%	0.00%	0.00%	100.00%	100%	100%
22	أقسام الأسم تبعاً للفظ آخره	46.96%	0.00%	0.00%	93.31%	100%	97.64%

5 الخاتمة

في هذا البحث عرضنا تجربة في توظيف معايير وأدوات التحليل الصرفي للغة العربية، بالإضافة إلى المعلومات اللغوية المستخرجة من كتب قواعد اللغة العربية التقليدية، لتدوير وإغناء المصادر اللغوية المتاحة بتحليل صرفيٍّ دقيق يحتوي على جميع الخصائص الصرفية للكلمة المحللة، وقد تمَّ استخدام مجموعة العناوين الصرفية (سلمى) كمعيار للتحليل الصرفي في هذه الدراسة، كما تمَّ استخدام المحلل الصرفي (سلمى) كأداة للتحليل الصرفي.

وقد تم تطبيق مبدأ تدوير المصادر المتاحة – أحد المبادئ الأساسية لهندسة البرمجيات – لتدوير وإغناء التحليل الصرفي لعينة من الذخيرة العربية القرآنية، وذلك بتحويل التحليل الصرفي آلياً من مجموعة العناوين الصرفية الخاصة بالذخيرة القرآنية إلى مجموعة العناوين (سلمى) ذات الخصائص الصرفية الدقيقة.

وقد تمت عملية التحويل أو التدوير لهذه العناوين الصرفية آلياً، وبرهنت النتائج عن قابلية تدوير وإغناء المصادر المتاحة بتحليل صرفي دقيق منتجة ذخيرة لغوية معنونة صرفياً بتحليل صرفي مفصل ودقيق.

المراجع

1. Sawalha, M. and E. Atwell) توظيف قواعد النحو والصرف في بناء محلل صرفي للغة العربية, *Adapting Language Grammar Rules for Building Morphological Analyzer for Arabic Language)*, in *Proceedings of the workshop of morphological analyzer experts for Arabic language, organized by Arab League Educational, Cultural and Scientific Organization (ALECSO), King Abdul-Aziz City of Technology (KACT) and Arabic Language Academy.* 2009: Damascus, Syria.
2. Sawalha, M. and E. Atwell, *Linguistically Informed and Corpus Informed Morphological Analysis of Arabic*, in *Proceedings of the 5th International Corpus Linguuistics Conference CL2009*. 2009: Liverpool, UK.
3. Sawalha, M. and E. Atwell, *Fine-Grain Morphological Analyzer and Part-of-Speech Tagger for Arabic Text*, in *Language Resource and Evaluation Conference LREC 2010* 2010: Valleta, Malta.
4. Dukes, K. and N. Habash, *Morphological Annotation of Quranic Arabic*, in *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. 2010, European Language Resources Association (ELRA): Valletta, Malta,19-21 May 2010.
5. Sawalha, M. and E. Atwell. *Comparative evaluation of Arabic language morphological analysers and stemmers*. in *Proceedings of COLING 2008 22nd International Conference on Computational Linguistics*. 2008.
6. Atwell, E., *Development of tag sets for part-of-speech tagging*, in *Corpus Linguistics: An International Handbook, Volume 1*, A. Ludeling and M. Kyto, Editors. 2008, Mouton de Gruyter. p. 501-526
7. Schmid, H. and F. Laws, *Estimation of Conditional Probabilities with Decision Trees and an Application to Fine-Grained POS Tagging*, in *COLING'08*. 2008: Manchester,UK.
8. Pauw, G.D. and G.-M.D. Schryver, *Improving the Computational Morphological Analysis of a Swahili Corpus for Lexicographic Purposes.* Lexikos 18 (AFRILEX-reeks/series 18: 2008), 2008: p. 303-318.
9. Jurafsky, D. and J.H. Martin, *Speech and Language Processing*. Second Edition ed. Prentice Hall Series in Artificial Intelligence, ed. S. Russell and P. Norvig. 2008, New Jersey: Prentice Hall. 1024.
10. Sawalha, M. and E. Atwell, *Constructing and Using Broad-Coverage Lexical Resource for Enhancing Morphological Analysis of Arabic*, in *Language Resource and Evaluation Conference LREC 2010*. 2010: Valleta, Malta.

الملخص باللغة الانجليزية

Title: Morphological Analysis of Modern Standard Arabic and Classical Arabic Text

Abstract. This paper shows the details of empirical study of applying standards and tools for Arabic morphological analysis, in addition to linguistic information extracted from long established traditional Arabic grammar books, to reuse and enriching existing resources with fine-grain morphological features information. SALMA Tag Set is used as standard in this work. And the tool used in this study is the SALMA Tagger [2, 3].

The reuse of existing components is an established principle in software engineering. We used the Quranic Arabic Corpus morphological annotation of a test text sample of chapter 29, consisting of about 1000 words. Then, an automated mapping methodology mapped the QAC morphological tags to SALMA morphological features tags.

The mapping between the QAC morphological tags and SALMA morphological features tags is done by following five steps procedure. This result proves that the reuse and enriching of existing resource with more detailed morphological features information is applicable and can provide a tagged corpora of fine grain analysis.

Keywords: Morphological analysis, Arabic language, The holy Qur’an, Reuse, SALMA - Tag Set, Gold standard for evaluating morphological analyzers of Arabic text

المصطلحات

Machine Learning	التعلم الآلي
Knowledge-base	الإعتمادية المعرفية
Reuse	تدوير
Software Engineering	هندسة البرمجيات
SALMA – Tag Set	مجموعة العناوين الصرفية (سلمى)
Lemmatizer	المُجَذِّع
Quranic Arabic Corpus	الذخيرة العربية القرآنية
Morphological Analyzer	المحلل الصرفي
Morpheme	مورفيم - مقطع
Part-of-Speech Tagging	العنونة الآلية بأقسام الكلام
Lemmatization	التجذيع
machine translators	المترجمات النَّصية
speech recognition	التعرف على الكلام
Ambiguity	الغموض
Productive	إنتاجية
Gold Standard	المعيار الذهبي
Accuracy	الدقة
Precision	الضبط
Recall	التذكر